3-15-2019

Examining MEDLINE Search Query Reproducibility and Resulting Variation in Search Results

C. Sean Burns
University of Kentucky, sean.burns@uky.edu

Robert M. Shapiro II
University of Kentucky, shapiro.rm@uky.edu

Tyler Nix
University of Michigan - Ann Arbor

Jeffrey T. Huber
University of Kentucky, jeffrey.huber@uky.edu

Digital Object Identifier (DOI)
https://doi.org/10.21900/iconf.2019.103369

Examining MEDLINE Search Query Reproducibility and Resulting Variation in Search Results

C. Sean Burns[1[0000-0001-8695-3643]], Robert M. Shapiro, II[1[0000-0003-4556-702X]], Tyler Nix[2[0000-0002-0503-386X]], and Jeffrey T. Huber[1[0000-0002-3317-0482]]

[1] University of Kentucky, Lexington KY 40506, USA

[2] University of Michigan, Ann Arbor MI 48109

Abstract. The MEDLINE database is publicly available through the National Library of Medicine's PubMed but the data file itself is also licensed to a number of vendors, who may offer their versions to institutional and other parties as part of a database platform. These vendors provide their own interface to the MEDLINE file and offer other technologies that attempt to make their version useful to subscribers. However, little is known about how vendor platforms ingest and interact with MEDLINE data files, nor how these changes influence the construction of search queries and the results they produce. This poster presents a longitudinal study of five MEDLINE databases involving 29 sets of logically and semantically consistent search queries (five search queries for each set). The goal is to understand whether it is possible to reproduce search queries by: a) analyzing search query syntax per database, and b) controlling for total search results. We also highlight the barriers to creating reproducible queries across MEDLINE databases.

Keywords: Information Storage, Information Retrieval, Search Queries, Medical Subject Headings (MeSH), MEDLINE

1 Introduction

Bibliographic databases are important to library and information scientists because these systems are designed to organize and retrieve information, such as records to books, journals, articles, and more [1]. Domain-oriented bibliographic databases, such as MEDLINE, exist and serve much the same role with respect to providing discovery and access to more specialized literature. MEDLINE provides a point of discovery for literature in the health, medical, life sciences, and related fields, and thus serves as an important resource for various audiences, including health care providers, health science librarians, life scientists and other researchers, as well as the general public [2].

The MEDLINE bibliographic data file is created and maintained by the National Library of Medicine (NLM) and publicly available online via PubMed. Other vendors license the MEDLINE file from the NLM and offer it on their platforms. MEDLINE is therefore also available via subscription from EBSCO*host*, ProQuest, Ovid, and Web of Science [3].

Although the MEDLINE file is presumably the same across these systems, the interfaces to the file and the search technologies on these systems differ. For example, one of the unique characteristics of the MEDLINE file is its use of the MeSH thesaurus.

Database vendors may have different technologies that treat the way the MeSH tree is searched and these technologies may impact how descriptors that exist on multiple branches, and that contain unique narrower terms depending on those locations, are indexed and retrieved when, for example, those terms are exploded.

Although the MEDLINE file may be the same among these vendors, there is little current research on whether this is true and also on how search deviates across these systems, given the differences in storage and retrieval technologies and in the interfaces that these vendors provide [4][5]. Given these differences, we hypothesize that vendor-based differences 1) make creating reproducible queries across these systems difficult, 2) that these difficulties change over time and impact search results over time, and 3) that problems with producing reproducible queries will impact search results. Therefore, understanding how queries function across what is assumed to be the same MEDLINE file is important because search results may affect the provision of health practice, and different search results may lead to variations in beliefs about medical practice and knowledge.

Therefore the purpose of this project is to focus on reproducible queries across these systems [6][7], and to determine 1) whether it is possible to reproduce search queries created in PubMed/MEDLINE in other MEDLINE databases; 2) to understand how and why queries may or may not be reproduced; and 3) to understand obstacles in producing consistent results across systems.

The project is motivated for several reasons. Health practitioners rely on MEDLINE to acquire evidence-based clinical information [8]. Although health practitioners may have access to the publicly available PubMed/MEDLINE database, they may also have institutional access to other platforms and prefer and use those more often or instead of PubMed/MEDLINE. These user preferences may become problematic in multi-institutional collaborations or in communicating research and data collection methods. Second, health researchers rely on MEDLINE to gather literature for research projects, including systematic reviews, which are often intended to guide medical practice. However, discovering this literature may be a function of the different interfaces that provide access to MEDLINE, among other factors, and thus controlling for search query syntax, or better understanding limitations of controlling queries across systems, may prove beneficial to literature discovery and gathering. There are also implications related to providing bibliographic instruction, whether that involves teaching future health care practitioners or future information professionals how to use MEDLINE.

2 Method

To answer our hypotheses, we are conducting a longitudinal study of five MEDLINE databases accessed through PubMed, ProQuest, EBSCO*host*, Ovid, and Web of Science. Since the MEDLINE file is updated daily in PubMed, and at unknown intervals in the other platforms, a longitudinal study will provide a more complete picture than a study based on a single date of data collection.

We created 29 sets of search queries with each set containing a query for each database mentioned above for a total of 145 total search queries. Each set is designed

to test a specific aspect of the search syntax among all five systems and to be logically and semantically consistent with the others in the respective sets. The search sets include basic keyword searches, searches against MeSH headings on single and multiple branches, searches with MeSH headings that explode, searches using different Boolean switches, searches with constraints on publication dates, limited to specific journal titles, and that combine some of the above. Final analysis of longitudinal results will include thematically classifying the obstacles presented by the different search syntax within the sets and by controlling for retrieved record counts. Table 1 reports an example search set and results.

Table 1. Example query set and results of MEDLINE databases. Search query is for *neoplasms* as a MeSH term and results are limited by publication dates 1950-2015.

Search #04	Search	Sept 2018
PubMed	"neoplasms"[MH:NOEXP] AND 1950:2015[DP]	349853
ProQuest	MESH.EXACT("neoplasms") AND YR(1950-2015)	347182
EBSCOhost	MH("neoplasms") AND YR 1950-2015	347195
Web of Science	MH=("neoplasms") AND PY=(1950-2015)	347173
Ovid	1. neoplasms.SH 2. limit 1 to YR=1950-2015	347184

The searches are conducted by two of the authors at two different institutions since no single author has access to all five of the MEDLINE databases. A pilot test was conducted in August 2018 in order to test the sets of queries. Data collection began in September 2018 and will continue monthly through August 2019.

3 Preliminary Findings

Data collection is ongoing, but we can report initial conclusions that relate to understanding obstacles in producing consistent results across systems. First, the documentation for the systems is poorly described and thus creating queries that are logically and semantically consistent with the others in a set is difficult. Since each of the vendors provides access to MEDLINE as well as other databases, we found that one obstacle to creating reproducible, or logically consistent queries, involves how the MEDLINE database inherits search technologies from the vendor. For example, field tags in all MEDLINE systems are dictated by the structure of the MEDLINE records, but some search operators are inherited from the vendor, such as ProQuest's EXACT operator, which may be used to control exploding MeSH terms but also inherits other uses since the operator is ProQuest specific and not MEDLINE specific. Field names, such as Publication Date, can be utilized by multiple vendors, but can have substantially different interpretations. In Web of Science, for instance, publication date information is determined by the Source field and does not support date ranges, whereas in PubMed, the Publication Date field represents the date that records were made public in Entrez and ranges are supported.

Second, MEDLINE's main characteristic is its use of the MeSH thesaurus. It is clear that PubMed applies search technologies that take advantage of the tree structure of the thesaurus, but other databases treat MeSH searches primarily as field searches, and this makes creating queries that explode MeSH terms difficult, especially if those MeSH headings exist on more than one branch of the MeSH tree. Third, even after controlling for publication date ranges and limiting results to publications that should be fairly fixed in the bibliographic record (1950-2015), we have found that the five databases do not agree by as much as several hundred records for some searches.

4 Discussion

Preliminary findings indicate that the old saying that "MEDLINE is MEDLINE is MEDLINE" is not currently accurate, and that we should be cautious about the queries we construct in MEDLINE systems. Furthermore, it is important to know the types of obstacles that users face when creating reproducible queries among MEDLINE systems, and that even when queries are logically consistent with others, search results may still vary. Some of the results may vary if the vendors update their MEDLINE files at different intervals, but this does not explain why results vary for queries that limit results to defined years. Given that health care providers, health science librarians, and scientists and researches depend on MEDLINE databases, this research should prove useful in clarifying how queries can be controlled, how they influence retrieval sets, and how this might influence data collection for research designs like systematic reviews or the collection of information for evidence-based medicine.

References

1. Palmer, C. L., Cragin, M. H. Scholarship and disciplinary practices. Annual Review of Information Science and Technology, 42(1), 163-212 (2008).
2. Medical Library Association. Role of expert searching in health sciences libraries. Journal of the Medical Library Association, 93, 42 (2005).
3. Bethel, A., Rogers, M. A checklist to assess database-hosting platforms for designing and running searches for systematic reviews. Health Information & Libraries Journal, 31, 46–53 (2014).
4. Hallet, K. S. Separate but equal? A system comparison study of MEDLINE's controlled vocabulary MeSH. Bulletin of the Medical Library Association 86(4), 491-495 (1998).
5. Parker, S. MEDLINE: Comparative review. Charleston Advisor 1(3), 5–10 (2000).
6. Peng, R. D. Reproducible research and biostatistics. Biostatistics 10 2009.
7. Goodman, S. N., Fanelli, D., Ioannidis, J. P. A. What does research reproducibility mean? Science Translational Medicine 8 (2016).
8. Kash, M. J. Teaching evidence-based medicine in the era of point-of-care databases: The case of the giant bladder stone. Medical Reference Service Quarterly 35(2), 230–236 (2016).